Insurance Claims- Fraud Detection with Machine Learning Technique

BY- Md Hafizur Rahman

What Does Fraudulent Claim Mean?

A fraudulent claim, in the context of insurance, is a claim based on a misrepresentation of facts with the intention of wrongfully gaining insurance benefits. A fraudulent claim is also known as a false insurance claim.



[This Photo](https://www.quoteinspector.com/15-of-our-best-free-auto-insurance-images/) by Unknown Author is licensed under [CC BY-ND](https://creativecommons.org/licenses/by-nd/3.0/)

A fraudulent claim may be classified as hard and soft fraud. The former is when a claimant deliberately plans, invents, or creates a loss covered by the insurance policy, while the latter occurs when a claimant exaggerates a legitimate claim or gives false information to obtain bigger gains. Making fraudulent claims is a crime, no matter what the outcome may be. Because they make up a significant portion of the total claims filed, losses for insurers are estimated to be in the billions every year.

**Problem Statement:**

**Business case:**  
Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

In this project, we are provided a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

In this example, we will be working with some auto insurance data to demonstrate how we can create a predictive model that predicts if an insurance claim is fraudulent or not.

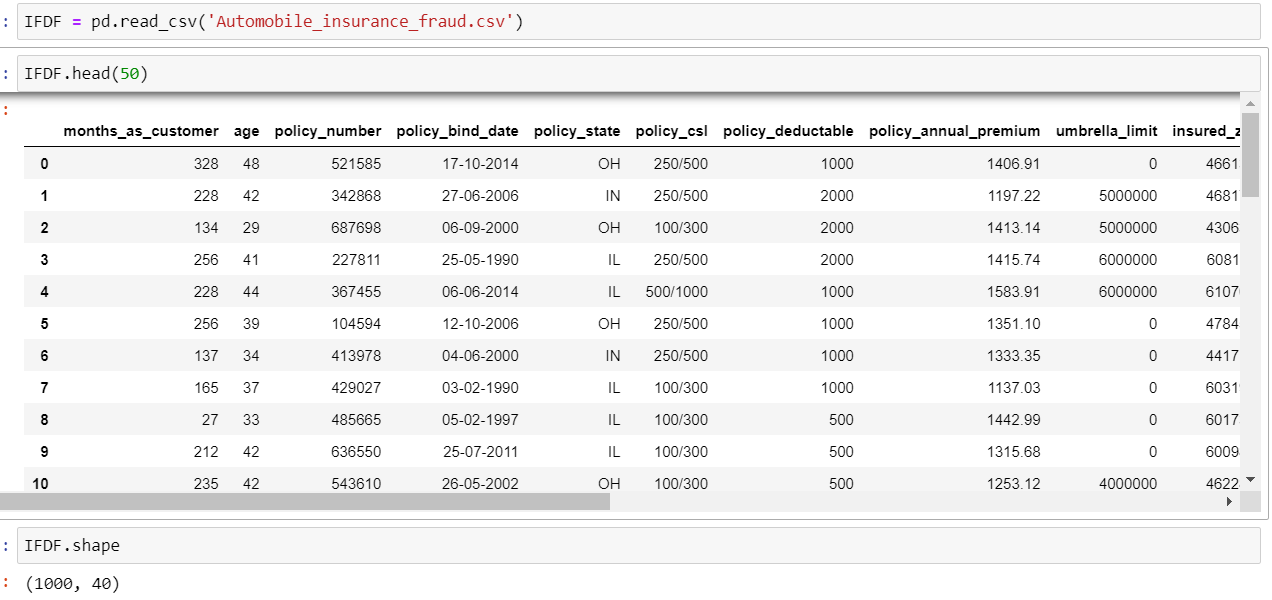
**Analytical Approach to Build the Model:**

In this project, a dataset was provided with the details of the insurance policy along with the customer details, as well as details of the accident on the basis of which the claims have been made.

The Dataset was first cleaned, the various feature columns were analysed, then with feature engineering and based on strength of correlation and ANOVA f-score values, the feature columns were selected that would best predict the Target variable, to train and test machine learning models.

The auto insurance dataset was worked with to build a predictive model that best predicts if an insurance claim is fraudulent or not. Several models were trained and fitted with a part of the dataset and then tested with a different part of the dataset. The model that performed the best with the best confusion matrix performance, f1 score, ROC-AUC score and cross validation performance was then selected and tuned further with hyper parameter tuning techniques.

**Dataset Description:**

****

**The given dataset consists of 1000 rows and 40 columns**.

**The Independent Feature columns are:**

**months\_as\_customer:** Number of months for which the person has been a customer

**age:**  Age of Customer

**policy\_number**: Identification number of policies

**policy\_bind\_date**: Time period between effective date of coverage and policy issuance.

**policy\_state**: State where policy is active

**policy\_csl:**  Policy Combined single limit

**policy\_deductable:** Amount paid before the insurance company starts paying up.

**Policy\_annual\_premium**: The total amount of premium paid annually

**Umbrella\_limit:** Provides excess limits and gives additional excess coverage

**Insured\_zip:** Zip Code of the Insured address

**insured\_sex :** Gender

**Insured\_education\_level:** Education Background of Insured

**Insured\_occupation:** Occupation of Insured

**Insured\_hobbies:** Hobbies of the Insured

**Insured\_relationship:** Relationship of the Insured

**Capital-gains:** Capital Gains made from insurance

**Capital-loss:** Capital Loss incurred

**Incident\_date:** Date on which Incident Occurred

**incident\_type:** Type of Incident

**Collision\_type:** Type of collision

**incident\_severity:** Severity of Incident

**Authorities\_contacted:** Whether authorities were contacted

**Incident\_state:** State where incident occurred

**incident\_city:** City where incident occurred

**incident\_location:** Location of incident

**Incident\_hour\_of\_the\_day:** Time of the day when incident occurred

**number\_of\_vehicles\_involved:** Number of vehicles involved in incident.

**property\_damage:** Whether there was property damage or not

**Bodily\_injuries:** Severity of bodily injuries

**witnesses:** Number of Witnesses

**Police\_report\_available:** Whether police reports are available

**Total\_claim\_amount:** Total amount of claim

**Injury\_claim:** Injury Claim amount

**Property\_claim:** Property Claim amount

**vehicle\_claim:** Vehicle Claim amount

**Auto\_make:** Make of Vehicle

**Auto\_model:** Model of Vehicle

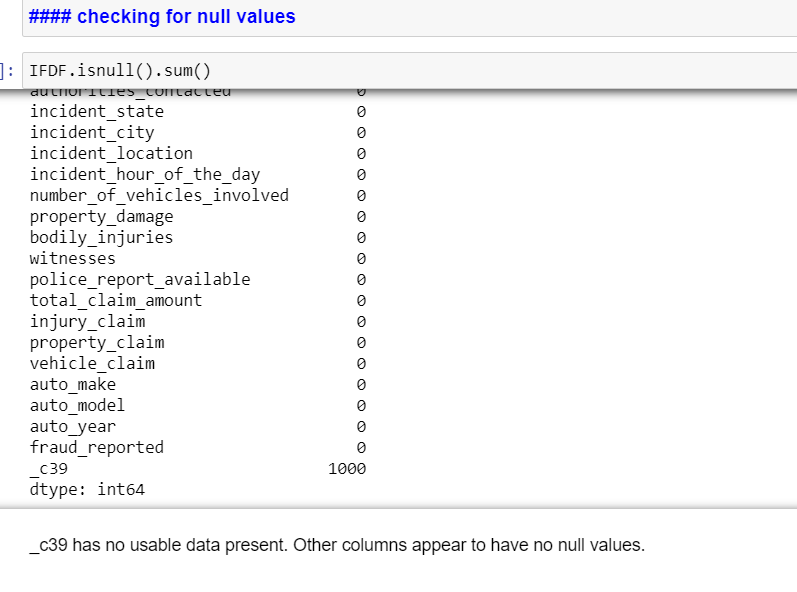
**Auto\_year:** Year of Vehicle Manufacture

**The Target Variable column:**

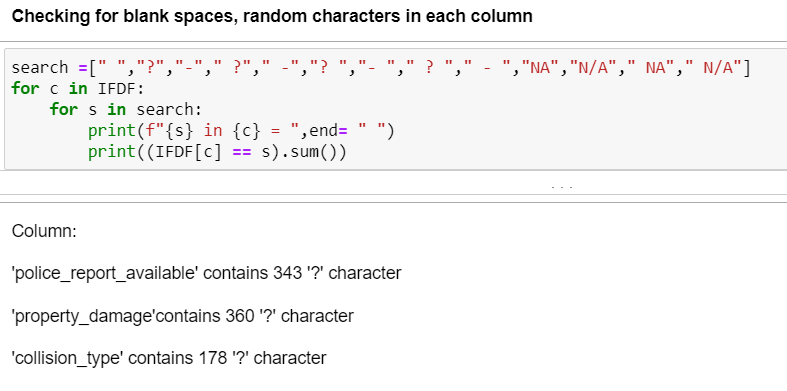
**Fraud\_reported:** Whether fraud was reported as Yes or No

**Looking For Null Values:**

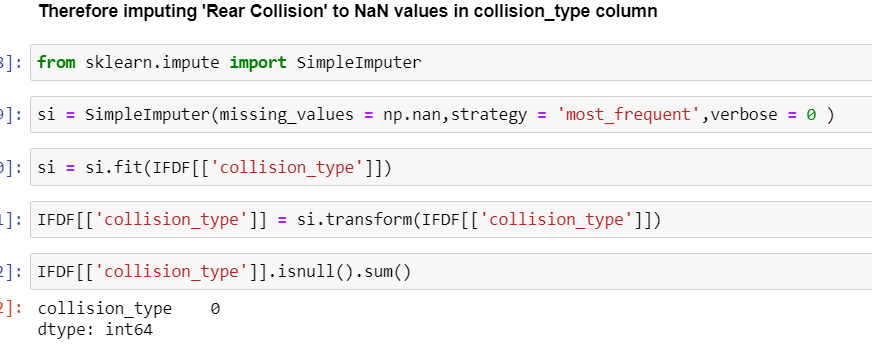
Upon inspecting all the columns in the data frame, it is observed that column \_c39 has no usable data present as they were all NaN values. Other columns appear to have no NaN values.



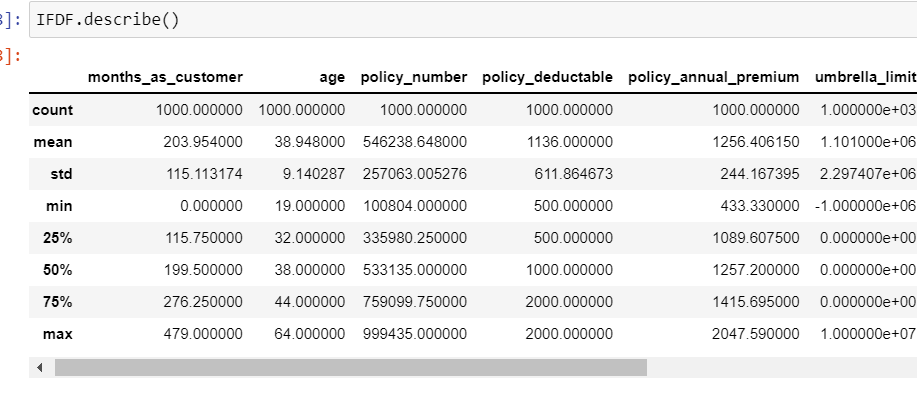
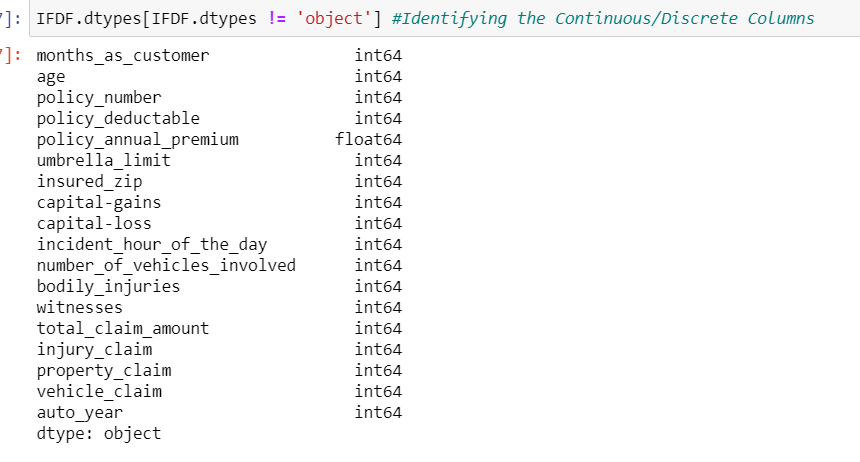
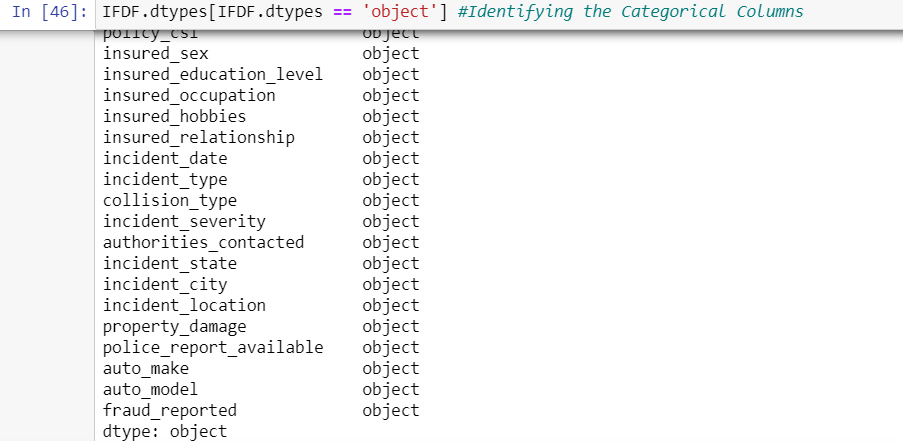
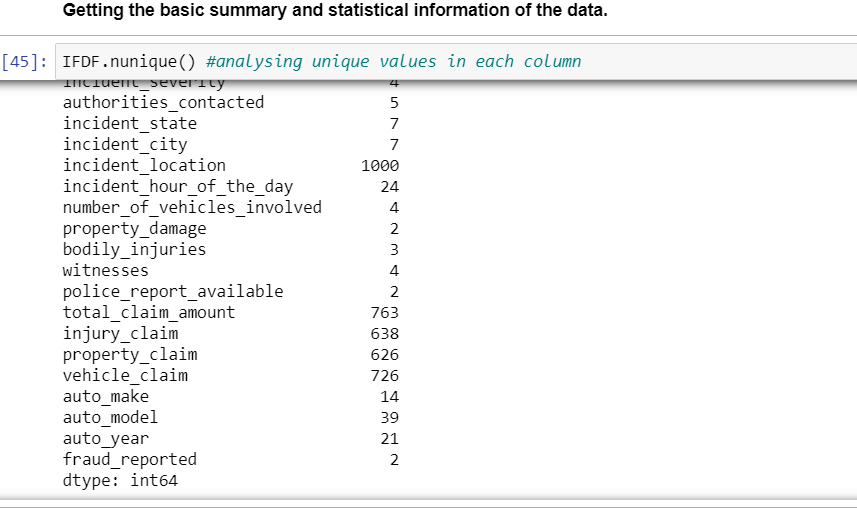
Further Cleaning the Data Set for insignificant values:



These values will be converted into NaN values which will be processed using imputation process.



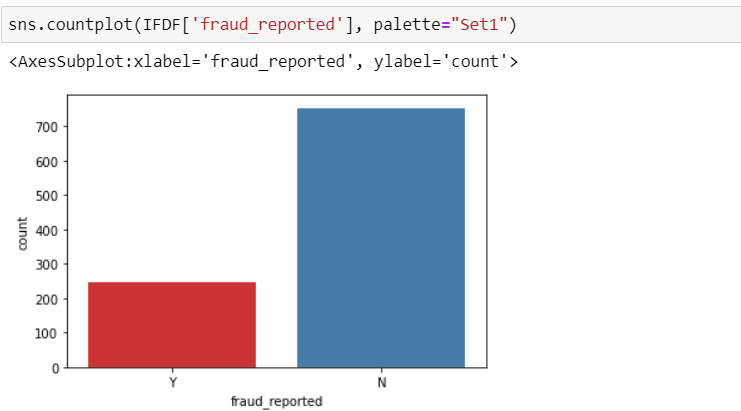
**After imputing the values, it is observed that no more null values remain in the dataset.**

**Performing EDA:**

Difference in mean and 50% and considerable difference in 75% and max of columns months\_as\_customer, policy\_annual\_premium, capital-gains,total\_claim\_amount, injury\_claim and property\_claim suggests skewness in respective data distributions and presence of outliers.

**Analysing Target Column:**

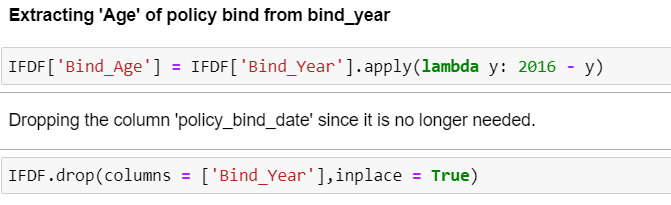
****

****

**Loking At the provided dataset we found that there are many columns from where the Exact Age is needed to be found:**

We use “Lambda Technique” to extract the Age Value of the relevant columns and later added the new columns to the dataset and removed the non-required columns.



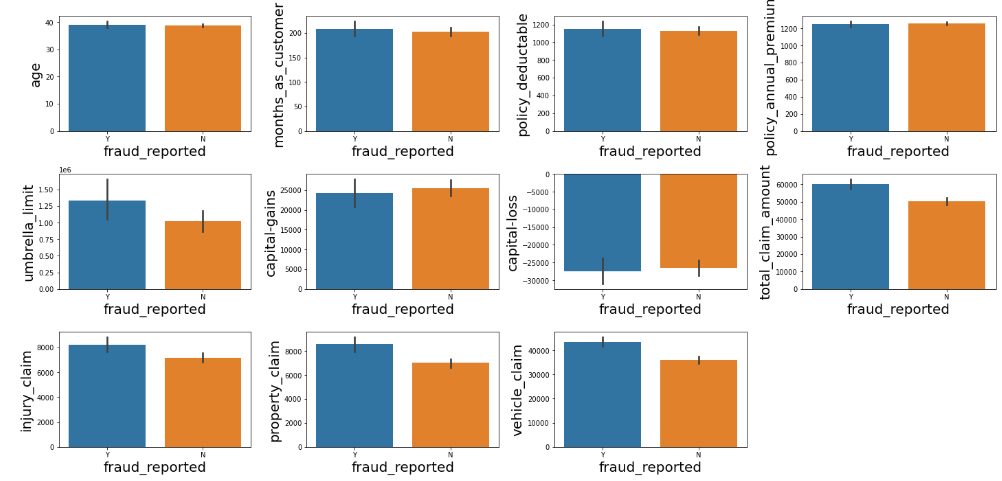


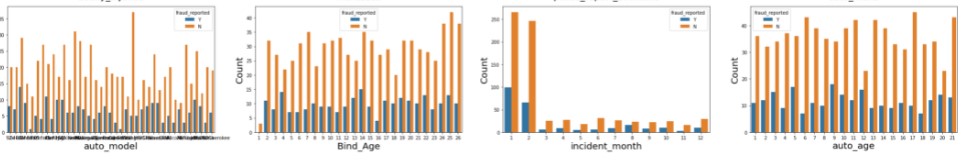
**Analysing rest of the feature columns:**

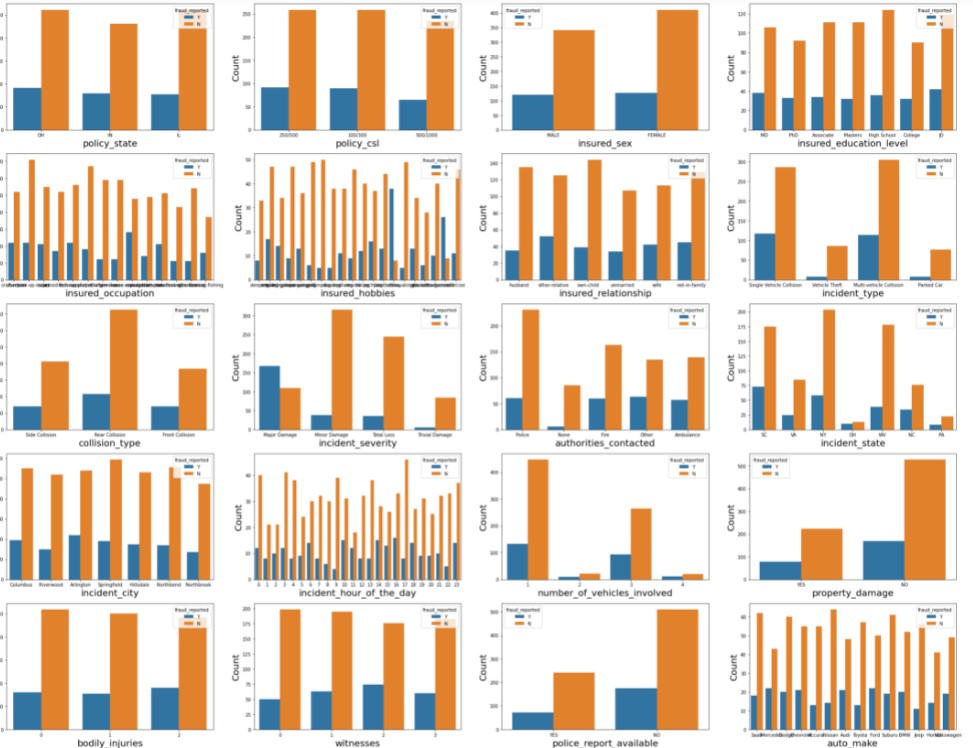
**Upon analysing the rest of the Feature Columns, following observations are made:**

* Majority of the cases are Multi-vehicle Collision and Single Vehicle Collision and are Rear Collisions.
* Most incidents take place between January and February.
* Minor Damage is most common followed by Major Damage and Total loss.
* Most common authorities contacted were the Police followed by Fire force.
* Most of the incidents occurred in states: SC, NY and WV
* Most incidents were reported from Columbus, Arlington, Springfield
* Majority reported no property damage.
* There are no police reports available for most cases.
* Most reports belong to models RAM, A3, Wrangler, Neon auto Models

### **Interpreting Relationship between Dependent Variable and Independent Variables:**







**Following observations can be made from above graphs:**

* 'age', 'months\_as\_customer','policy\_deductable','policy\_annual\_premium','capital-gains','capital-loss', don't seem to contribute to fraud probability.
* Higher the umbrella limit, more fraud claims are filed.
* Higher the total claim amount, more the fraud claims are filed.
* Higher the injury claim amount, more the fraud claims are filed.
* Higher the property claim amount, more the fraud claims are filed.
* Higher the vehicle claim amount, more the fraud claims are filed.
* policy state, policy csl, insured sex, authorities contacted, bodily injuries, incident city, witnesses don't seem to contribute to fraud probability.
* Education levels of JD and Highschool and MD contribute most to the fraud claims filed.
* Relationships - other relative and not in family contribute most to the fraud claims filed.
* Single vehicle collision and multi vehicle collision contribute most to the fraud claims filed.
* Incidents in states SC and NY contribute most to the fraud claims filed.
* fraud claims are more for 1 and 3 vehicles involved in accident
* fraud claims are more for rear collision in accident
* fraud claims are most for Major damage reported
* fraud claims are most for hours 10,14,16,18(office rush hours) and 23 of the day
* fraud claims are more when no property damage is reported
* fraud claims are more when no police report is available
* fraud claims are more during months 1(January) and 2(February).
* fraud claims are policy bind ages 2,4,13 and 14
* fraud claims are most for car age 3,5,9,12,20,21.
* Ram, A5, Jetta, ML350, Passat, F150, A3 have the highest fraud insurance claims, while 3 series, RSX, Camry have the lowest.
* Wrangler, Passat,95, Neon, Malibu, Grand Cherokee, auto\_model\_Ultima, Corolla, TL, Legacy has the highest legitimate claims.
* Mercedes, Dodge, Chevrolet, Audi, Ford, Volkswagen have the highest fraud insurance claims.
* Most fraud reports were filed by exec managerial, Transport moving and Craft repair.
* Most fraud claimants have chess and cross fit as hobbies

**Checking for Outliers in columns with continuous distribution:**

plt.figure(figsize=(10,5),facecolor='white')

plotnum=1

for col in IFDF[['vehicle\_claim','total\_claim\_amount','umbrella\_limit']]:

if plotnum<=3:

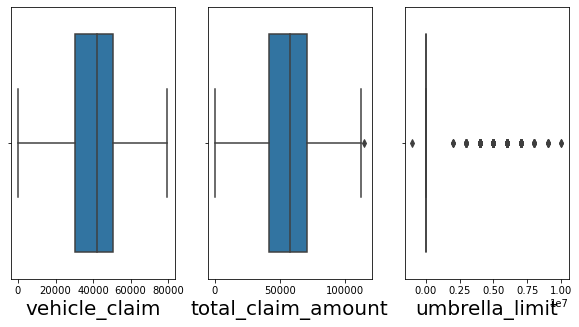
plt.subplot(1,3,plotnum)

sns.boxplot(IFDF[col])

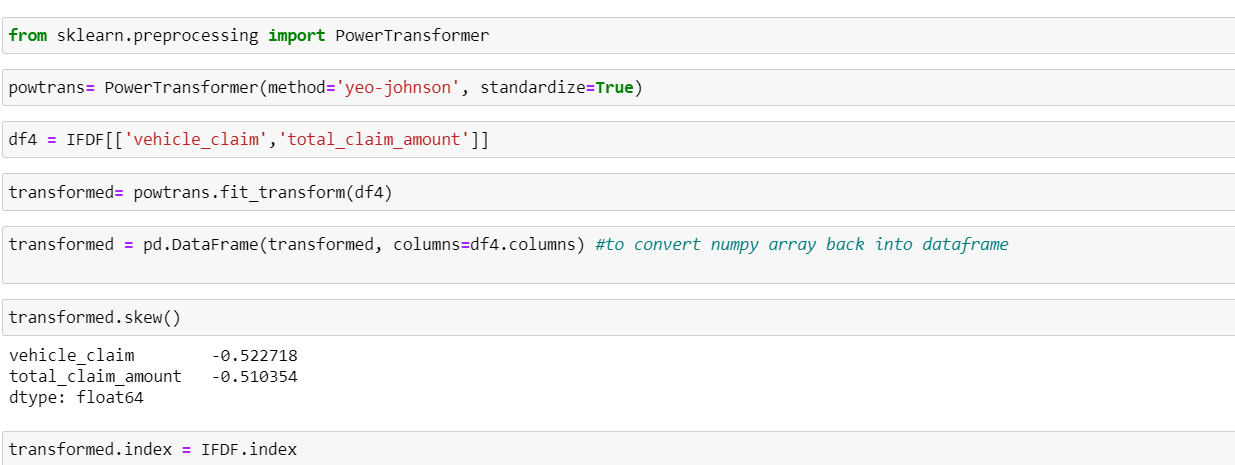
plt.xlabel(col,fontsize=20)

plotnum+=1

plt.show()



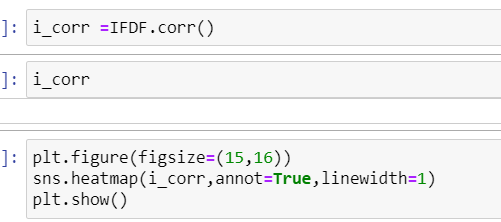
**Then we remove skewness using PowerTransform:**

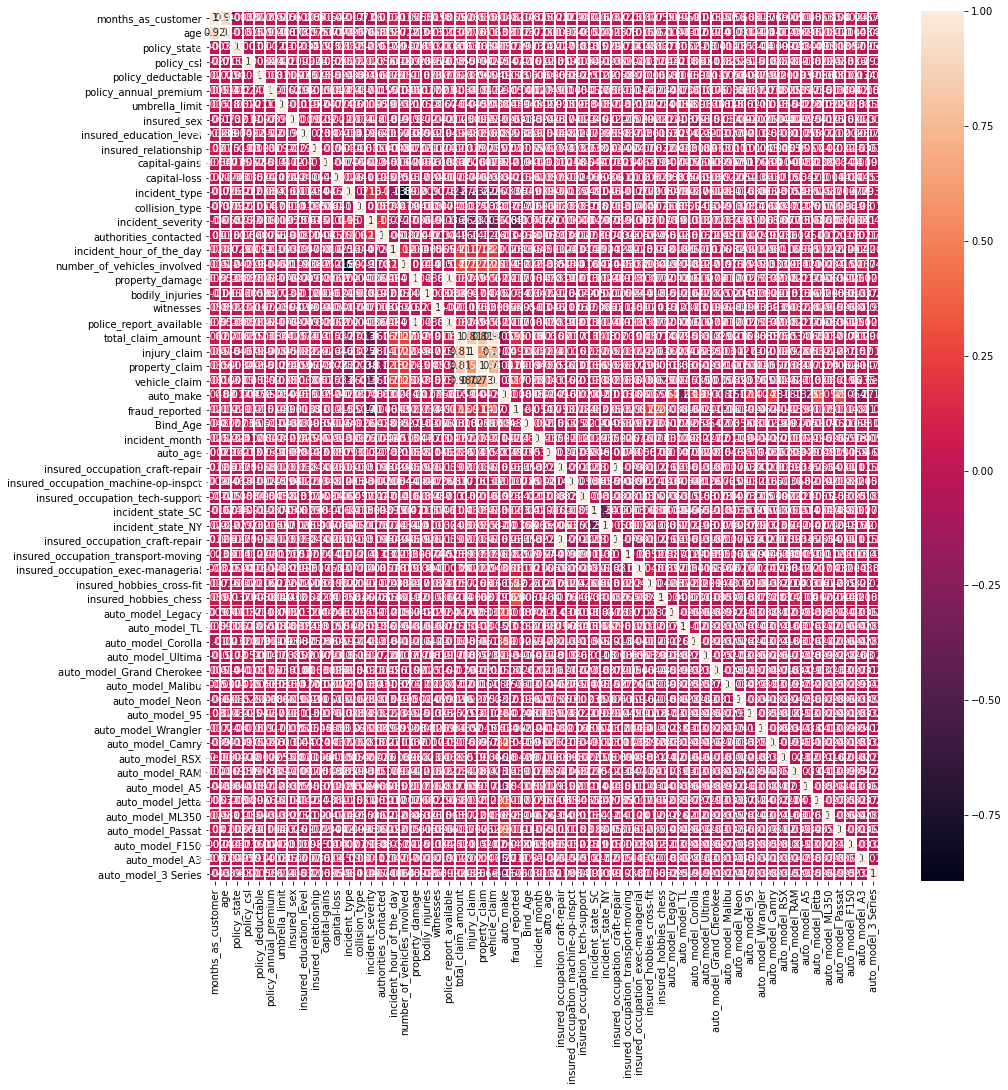
****

**Encoding the Categorical Columns in the Data Set:**

****

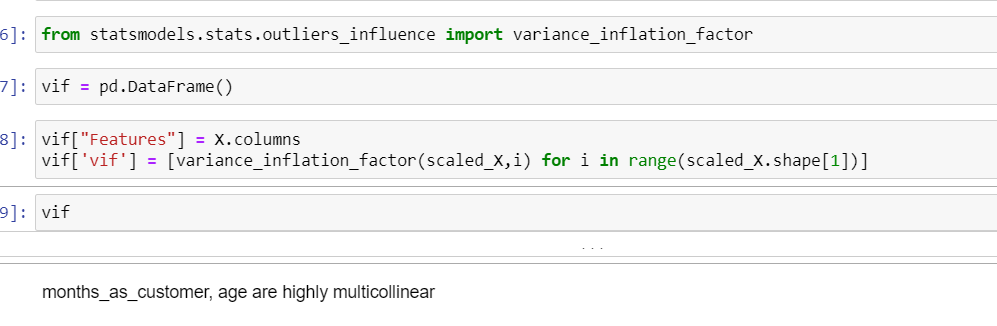
**Finding Correlations:**

****

****

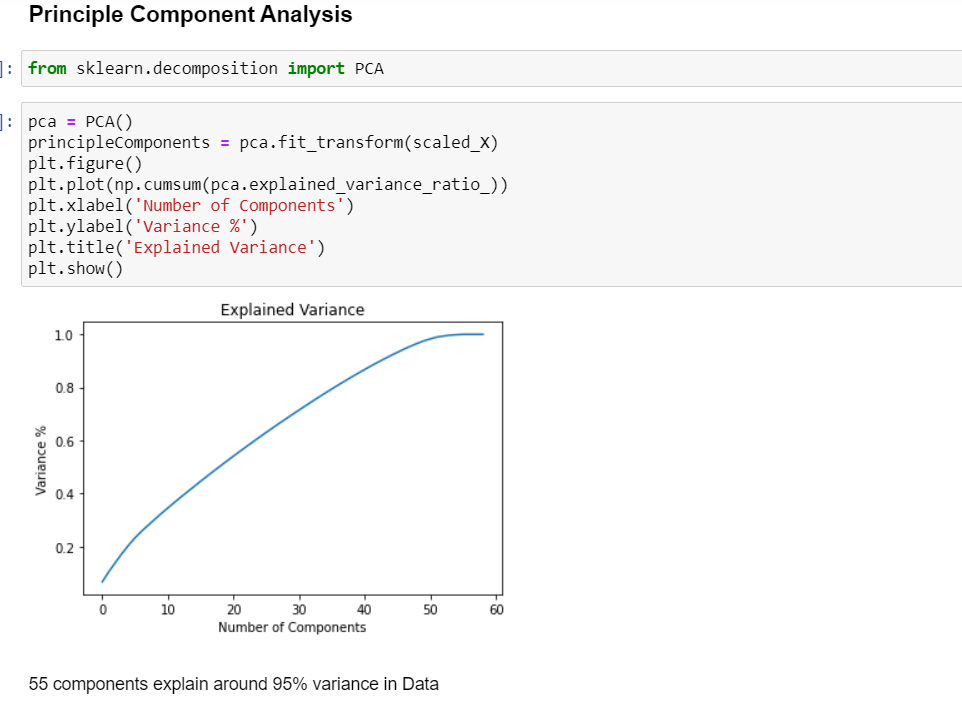
Incident\_severity has the highest negative correlation with fraud\_reported while, insured\_hobbies\_chess, insured\_hobbies\_cross-fit, vehicle\_claim, total\_claim\_amount, property\_claim have the highest positive correlation with fraud\_reported.

**Checking for Multicollinearity using Variance Inflation Factor:**



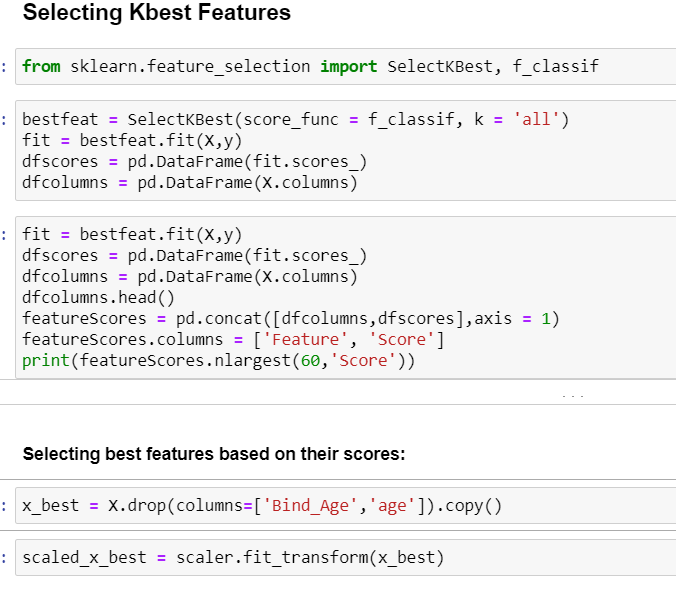
Variance inflation factor measures how much the variance of an independent variable is influenced / inflated, by its interaction/correlation with other independent variables.

**Performing Principal Component Analysis:**



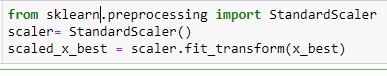
**Selecting KBest Feeature:**

Based on the respective ANOVA f-score values, the feature columns are selected that would best predict the Target variable, to train and test machine learning models.



Upon analysing the scores of each column, it is decided that the columns with the lowest scores will be dropped.

**Feature scaling:**

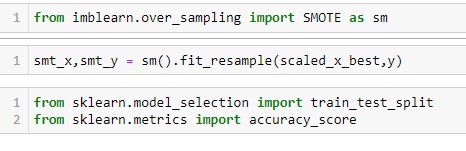
Scaling the values in the feature columns using StandardScaler inorder to normalize the range of data**.**

**Balancing out classes in the Label column using SMOTE technique.**

The classes in the target column are heavily imbalanced. Inorder to ensure that the precision and recall accuracies of the models for both of the classes are balanced, the classes of the target column need to be balanced.

SMOTE technique will be implemented to balance them out.

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.



**Analyzing Model Accuracies:**

**Logistic Regression Model Accuracy**

The trained Logistic Regression Model shows

F1 score of 0.87

Roc\_auc score of 0.8693

Cross validation score of 0.8466

Sensitivity (Recall for ‘Fraud’ cases) is 0.85 and Specificity (recall of non-fraud cases) is 0.88

Precision for ‘Fraud’ cases is 0.88 and Precision for non-fraud cases is 0.87

**Random Forest Classifier Model Accuracy**

The trained Random Forest Classifier Model shows

F1 score of 0.89

Roc\_auc score of 0.9057

Cross validation score of 0.8799

Sensitivity (Recall for ‘Fraud’ cases) is 0.89 and Specificity (recall of non-fraud cases) is 0.89

Precision for ‘Fraud’ cases is 0.89 and Precision for non-fraud cases is 0.89

**XGB Classifier Model Accuracy**

The trained XGB Classifier Model shows

F1 score of 0.90

Roc\_auc score of 0.9121

Cross validation score of 0.8799

Sensitivity (Recall for ‘Fraud’ cases) is 0.88 and Specificity (recall of non-fraud cases) is 0.91

Precision for ‘Fraud’ cases is 0.91 and Precision for non-fraud cases is 0.90

**AdaBoost Classifier Model Accuracy**

The trained AdaBoost Classifier Model shows

F1 score of 0.88

Roc\_auc score of 0.8886

Cross validation score of 0.8600

Sensitivity (Recall for ‘Fraud’ cases) is 0.88 and Specificity (Recall of non-fraud cases) is 0.88

Precision for ‘Fraud’ cases is 0.86 and Precision for non-fraud cases is 0.89

**SV Classifier Model Accuracy**

The trained SV Classifier Model shows

F1 score of 0.89

Roc\_auc score of 0.8971

Cross validation score of 0.8871

Sensitivity (Recall for ‘Fraud’ cases) is 0.88 and Specificity (Recall of non-fraud cases) is 0.92

Precision for ‘Fraud’ cases is 0.91 and Precision for non-fraud cases is 0.88

**K Nearest Neighbors Classifier Model Accuracy**

The trained K Nearest Neighbors Classifier Model shows

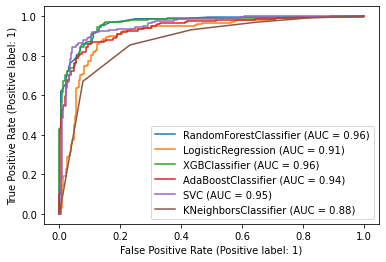
F1 score of 0.69

Roc\_auc score of 0.7276

Cross validation score of 0.7583

Sensitivity (Recall for ‘Fraud’ cases) is 0.93 and Specificity (Recall of non-fraud cases) is 0.57

Precision for ‘Fraud’ cases is 0.68 and Precision or non-fraud cases is 0.89

**ROC AUC curves:**

Based on the graph and roc\_auc\_scores,XGB Classifier is the best model for the dataset, with AUC = 0.96 and roc\_auc\_score = 0.8993.

**Hyper Parameter Tuning**

GridSearchCV was used for Hyper Parameter Tuning of the XGB Classifier model.

Based on the input parameter values and after fitting the train datasets,

The XGB Model was further tuned based on the parameter values yielded from GridsearchCV.

* The Tuned XGB Model displayed accuracy of 90.57%
* F1 score of 0.90
* Sensitivity (Recall for ‘Fraud’ cases) is 0.88 and Specificity (recall of non-fraud cases) is 0.91
* Precision for ‘Fraud’ cases is 0.89 and Precision for non-fraud cases is 0.91

**Concluding Remarks:**

In conclusion, XGB Classifier Model is able to correctly distinguish between Fraud claims and legitimate claims with high accuracy.

The dataset had very limited data which is problematic as models show greater stability when the dataset is of a good size. However, a large set of feature columns enabled selecting a smaller feature size that provides the best accuracy for the model and obtaining results in optimal time.